



Fraunhofer Institut
Experimentelles
Software Engineering

A Handbook for Statistical Data Analysis of Code Measurement Data

Authors:

Ester Palacios Rodriguez
Teade Punter

IESE-Report No. 140.03/E
Version 1.0,
September 30, 2003

A publication by Fraunhofer IESE

Fraunhofer IESE is an institute of the Fraunhofer Gesellschaft.

The institute transfers innovative software development techniques, methods and tools into industrial practice, assists companies in building software competencies customized to their needs, and helps them to establish a competitive market position.

Fraunhofer IESE is directed by
Prof. Dr. Dieter Rombach
Sauerwiesen 6
67661 Kaiserslautern

About this Document

This is a Handbook that is meant for IESE-employees that will use STATISTICA to conduct statistical analysis of M-System Data.

Table of Contents

1	Introduction	1
	Process Overview	1
2	Data analysis issues	3
	Goals and data set	3
	Data Analysis techniques and activities	4
	Macro for Statistical Analysis of M-System Data	6
3	Data analysis process	7
3.1	Reading the data files	7
3.2	Selection of the variables on behalf of the hypotheses	8
	Determine statistical inference	9
3.3	Select appropriate methods	11
3.3.1	Perform Descriptive Statistics & Outlier Analysis	12
	Example	13
3.3.2	Principal Component Analysis	17
	Introduction	17
	Model	18
	How apply PCA	19
3.3.3	Correlation of measures to size - Optional	20
3.3.4	Univariate Analysis	20
	Introduction	20
	Logistic Regression	21
	Assessing Fit	22
	Poisson Regression – Optional step	23
	Example: Logistic Regression	24
3.3.5	Multivariate Analysis.	25
	Introduction	25
	Logistic Regression	25
	Example: Logistic Regression	27
3.3.6	Evaluating estimation accuracy	28

3.3.7	Cross-Validation	29
	Introduction	29
	Examples: Cross-Validation	30
3.3.8	Model Application	31
4	Summary	32
5	References	32
6	Appendix A.1: Tool Usage	35
	Hardware & Software Requirement	35
7	Appendix A.2: Statistica Macro	36
	The following describes the installation of the Statistica macro.	36
	How to Start	36
8	Appendix B: Kappa	38
9	Appendix C: Validity threats	40
	Index	42

1 Introduction

Analysis for software measurement data is essential to assess the quality of software developed. Building prediction models for external attributes of systems or investigate particular relationships, e.g. size with some specific quality attribute, are able to capture information about past processes and use it to help make judgment about existing or future ones [5].

Data Analysis provides a means to make software measures comparable and determine the characteristics of metrics and relationships among them [5].

Process Overview

M-System is a tool suit for measurement of systems that calculates a set of measures (OO product measures) from structural properties for a software system. The process to apply M-system is presented below.

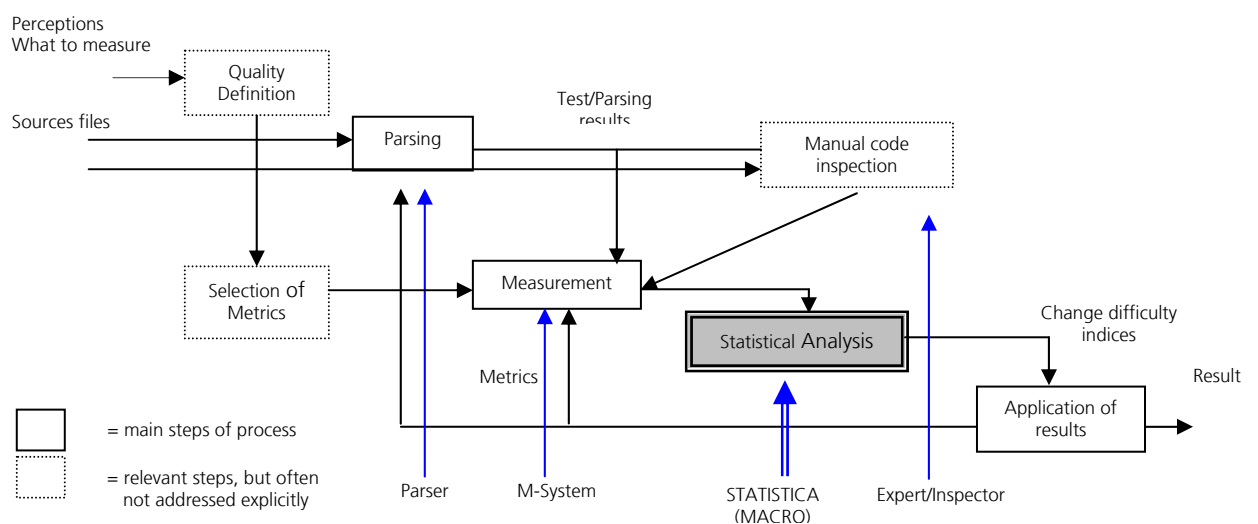


Figure 1: M-system process ([4], page 4)

In the earlier stages of the process (see figure1), quality definition and metrics specification activities yield the quality model and its associated metrics. Parsing of the source code and data retrieval of measurement, provide the necessary

outcomes (M-System data) for the next stage: **Statistical Analysis**, where the results will be analysed in an appropriate way.

This manual aims at describing the process of analysing statistically measurement data from M-System to investigate the relationships between product measures and external quality attributes, e.g. effort, fault-proneness, etc. A Tool Support for this process is provided as well.

The manual is organized as follows. Section 2 begins by introducing the process of analysing M-System Data that will be described in the later section. Section 3 presents the data analysis process as well as some examples with the macro developed for that purpose. STATISTICA [17] is used as Tool Support. Section 4 summarizes the results. Section 5 provides several resources about the techniques and Tools used for identical purpose. The handbook also includes an Appendix with a brief introduction, installation and usage of the Tool (Macro) using STATISTICA.

2 Data analysis issues

Goals and data set

The data analysis is the analysis of the provided data set according to the defined goals. Before the data analysis can start, the goals of the study have to be defined and the data should be collected.

The Goals of the study have to be defined during the phase quality definition (see figure 1). Goals are preferably defined according the GQM goal measurement template. When defining goals as well as specifying the metrics, also a set of hypotheses and the collection of variables, which will influence the hypotheses, will be defined.

The Data set result from the M-system. They are M-System Data: All measures to be analysed, are extracted from the system in study with the M-System tool. The results are stored in files that will be used later.

The results provided by M-System are stored in files according the following format.

sys	class	LCOM1	LCOM2	LCOM3	LCOM4	LCOM5	Coh	Co	LCC	TCC	ICH
./mmm2	OS	74	57	5	3	0.93706	0.12987	0.08333	0.75000	0.42857	16
./mmm2	UCA	3	3	3	3	1.00000	0.33333	-1.0000	0.00000	0.00000	4
./mmm2	LEC	2	1	2	2	1.46000	0.02666	-0.5000	0.33333	0.33333	0
./mmm2	RMI	0	0	1	1	0.00000	1.00000	1.00000	1.00000	0	
./mmm2	Sed	1	1	2	1	2.00000	0.00000	1.00000	0.00000	0.00000	4
./mmm2	Loq	2415	2345	44	36	0.99693	0.01710	0.00227	0.08134	0.05754	46

Figure 2: Output file format for cohesion metric. Filename: name.coh.dump

These files will be used as input files for the Macro that implements the statistical techniques described in the following sections. The first two columns describe the system and the classes, to which the measures belong. The rest of the columns are the measures obtained from M-System.

Data Analysis techniques and activities

Data analysis: Depending on the purpose or objective of the study, it can investigate particular characteristics of the data and explore relationships between the measures. Several statistical techniques (see figure 3) are presented in this manual for that purpose (software metrics analysis).

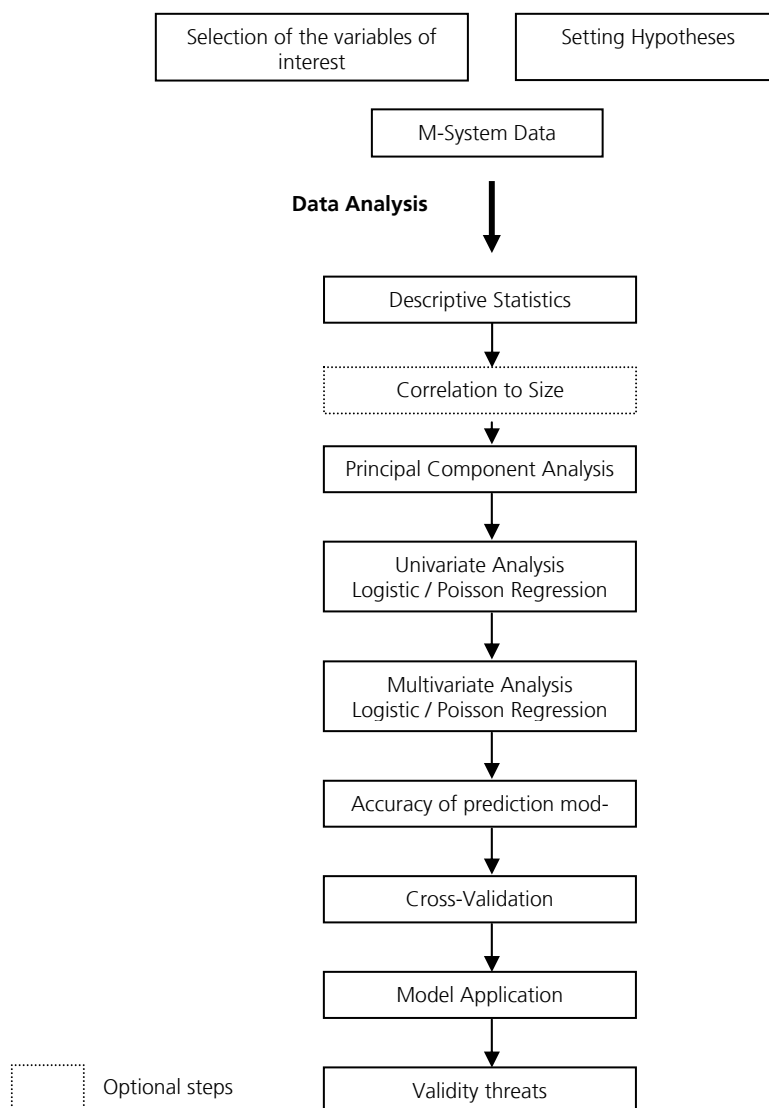


Figure 3. Process of analysing of M-System data.

The figure 3 shows the whole process for analysing M-System Data. This process is proposed in [1] [2][3] like a precise, complete and repeatable analysis pro-

cedure, which, when followed in replicated studies, enables comparisons to be made across them.

Text Boxes show the options for analysing according to criteria. The order should be, from top to bottom (see figure 3): descriptive statistics, then, optionally correlation with size, then PCA, univariate analysis, multivariate analysis, fit, cross-validation and model application.

An overview of the process is presented below

Descriptive Statistics: give a first idea about how the data are distributed. These results can draw conclusions about the selection of technique to be used.

Correlation to Size: Investigate the relationship between design measures and the size of the classes can help to find out if the measure (coupling, cohesion, etc) is measuring size ([1],[2]). This can help to understand its relationship with the quality attributes in study (effort, fault-proneness, etc).

Principal Component Analysis (PCA): Is a statistical technique that is used to identify the number of dimensions of the product measures and reduce the number of them to be used in later techniques.

Univariate Analysis: Investigate each individual measure against the quality attribute to be study (dependent variable), is performed with Univariate Analysis.

Multivariate Analysis: Investigate, how well multiple measures, when they are used in combination, predict the dependent variable is performed with Multivariate Analysis.

Fit: Evaluate the accuracy of the predictions models and give an estimation about how well is the solution

Cross-Validation: Estimate the accuracy when using the model on a different data set

Model Application: Usage of the prediction models built in practice (during software development process) and evaluation of its performance.

Validity threats: It is desirable to validate the results and the possible risks that are taking in account during the study.

Macro for Statistical Analysis of M-System Data

The screenshot shows a software window titled "Statistical Analysis of M-System Data". It contains several functional areas:

- Descriptive Statistics & Outliers (1):** Includes buttons for "Summary" and "Outliers", and a text input field "N >" with the value "6".
- Principal Component Analysis (2):** Includes a "PCA Run" button and a checked checkbox labeled "Minimum eigenvalue = 1".
- Univariate Analysis (3):** Includes a "Logistic Regression" button and a text input field "alpha" with the value "0.05".
- Multivariate Analysis (4):** Includes a "Logistic Regression" button.
- Cross-Validation (5):** Includes a "Run C-V" button.

On the right side of the window, there are two additional buttons: "Load files" and "Cancel".

Figure 4. Form for data analysis

To support the data analysis process, a macro has been implemented in STATISTICA. Figure 4 presents the control panel for this macro and shows the functionality for supporting the data analysis process. Details about Installation and Usage can also be found in the appendix described in section 7

The application (see figure 4) provides all functionality and gives support to complete process (see figure 3).

Note that although some steps in the process (also named with numbers in the form) can be applied without any previous treatment of the data, for the steps 4 (Multivariate Analysis) and 5 (Cross-Validation) are necessary results of the earlier steps and some kind of analysis.

3 Data analysis process

3.1 Reading the data files

Previous chapter has already denoted that the data set (from M-system) is the starting point for the analysis. The following provides the process how to read the files into the STATISTICA Macro.

1) Click the button "Load Files".

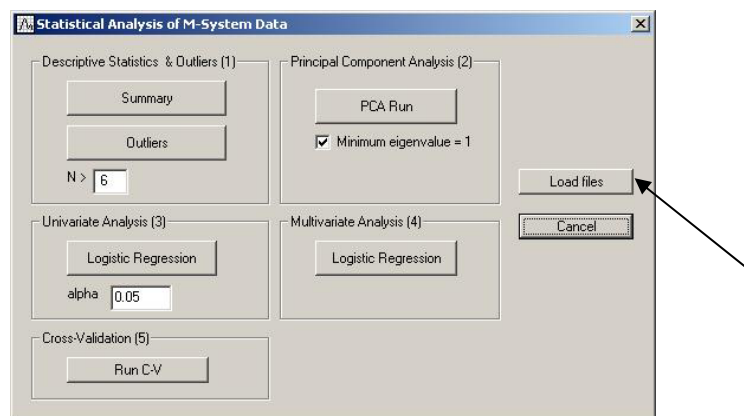


Figure 5.

Control Panel for the data analysis

2) Select the files involved in the analysis.

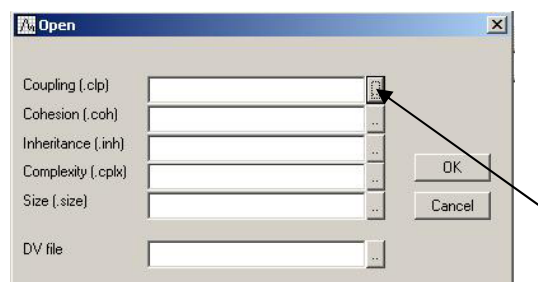


Figure 6.

Open data file dialog

3) Select a file and click ok.

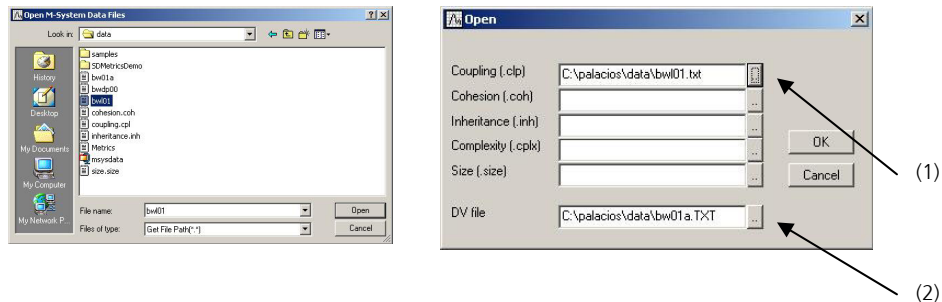


Figure 7. Open Dialog forms.

At least there must be loaded two files (they can be the same). One (1) in figure 7) for independent variables and (2) for the dependent variable.

Click Ok.

4) To open each of the files selected (see figure 7), the following two dialogs has to be responded.

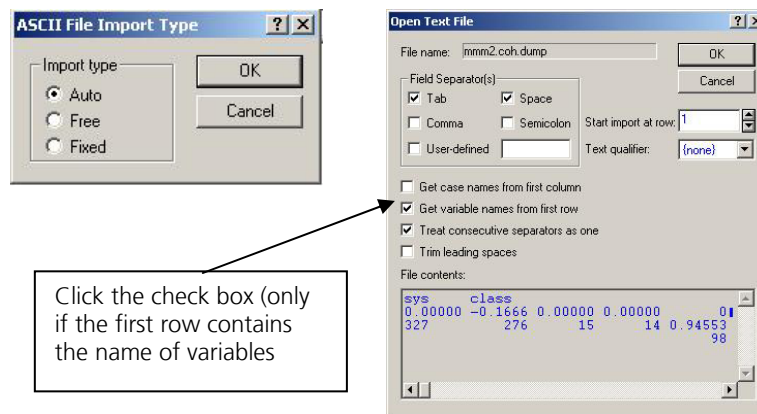


Figure 8. Options. Open data file dialogs.

The control panel (see figure 5) is displayed again on the screen.

3.2 Selection of the variables on behalf of the hypotheses

When investigate the characteristics of the data or when the relationships between measures and other factors (attributes) are to be explored, the goals of the study should be reflected in the hypotheses and the selection of variables. **Hypotheses** are statements that explain the behaviour that is to be explored.

They are affirmations of the relationships among variables, which are investigated. An example of hypothesis is presented in example 1.

Example 1: *A class with low cohesion is more likely to be fault-prone than a class with high cohesion. Low cohesion indicates inappropriate design, which is likely to be more fault-prone [1].*

Note: the hypotheses should be set during the phase Quality definition (see figure 1). Once the set of hypotheses is stated, the set of variables of interest that affect that set of hypotheses must be defined. A variable is some characteristic or property that differs in value from one case to another. **Variables** can be divided in two types: **independent variables**, also called predictors and **dependent variables** that are the consequences of the dependent variable.

Example 2: *In example 1, cohesion can be taken as the independent variable and fault-proneness as the dependent variable. Which relationship e.g. might be hypothesized as that less cohesion will result into higher fault-proneness.*

In this case, the independent variables are the M-System data collected in the previous phase. That is, the different measures (cohesion, inheritance, etc) extracted from the system. It will be the task of the User to decide what measures of interest and quality attribute as dependent variable will take part in the study. Usually, quality attributes as effort, cost, fault-proneness, etc, are chosen as dependent variables for the study.

Examples of dependent variables [11]:

- Fault proneness
- Development effort
- Test effort
- Rework effort
- Reusability
- Maintainability

Investigate the relationships among measures and how they influence on the dependent variable will be the main data analysis purpose.

Determine statistical inference

The logic of statistical inferences is based on two possible outcomes [5]:

- 1) Null Hypothesis H_0 : is a statement about the state of the world to be rejected

Example 3: From example 1, a class with low cohesion is **not** more likely to be fault-prone than a class with high cohesion.

- 2) Alternative Hypothesis H_1 : It is a statement about the state of the world to be accepted

Example 4: From example 1, a class with low cohesion is more likely to be fault-prone than a class with high cohesion.

The null as well as the alternative hypotheses should be stated clearly so that it can be evaluated at the end of the study [10].

To accept the null hypothesis, the statistical significance (α -level) is evaluated. Alpha is a parameter to indicate the probability of rejecting the statistical hypothesis tested (H_0) when in fact, that hypothesis is true. Usually, α is set at 0.05 (5%) or 0.01(1%). That is, if the obtained statistic (p) is between assumed 0.95 or 0.99, then the results of statistical test are significant and the samples are drawn from a population in which the hypothesis is true, otherwise (the obtained statistic is lower than 0.95 or 0.99), the statistical test are not significant (are due to chance)

Figure 5 shows the decision table for testing. β is a parameter to indicate the probability of failing to reject the hypothesis tested when that hypothesis is false and a specific alternative hypothesis is true. For a given test, the value of beta is determined by the previously elected value of alpha.

- Accepting the null hypothesis when is false: Type II error.
- Rejecting the null hypothesis when is True: Type I error.

Therefore, Alpha is the probability of committing the Type I error. The probability of Type II error is equal to Beta (also called the power of the statistical test). If test results indicates to reject H_1 , then H_0 is accepted, but if test does not reject H_1 , only it can be stated that there is not ground to reject H_0 .

	H_0 True	H_0 False
Reject H_0	Type I error: α	No error: $1-\beta$
Accept H_0	No error: $1-\alpha$	Type II error: β

Figure 9.

Decision table for testing.

3.3 Select appropriate methods

Many statistical techniques for data analysis are presented in the literature. This Manual focuses in those that are of interest to investigate software measurement data.

Since the type of distribution determines the statistical technique to be applied, the nature and distribution of the data must be explored in advance. In this manual, a non-normal distribution is assumed. That is because in many cases of study existing in the literature ([1],[2]), it is shown that data are non-normal distributed.

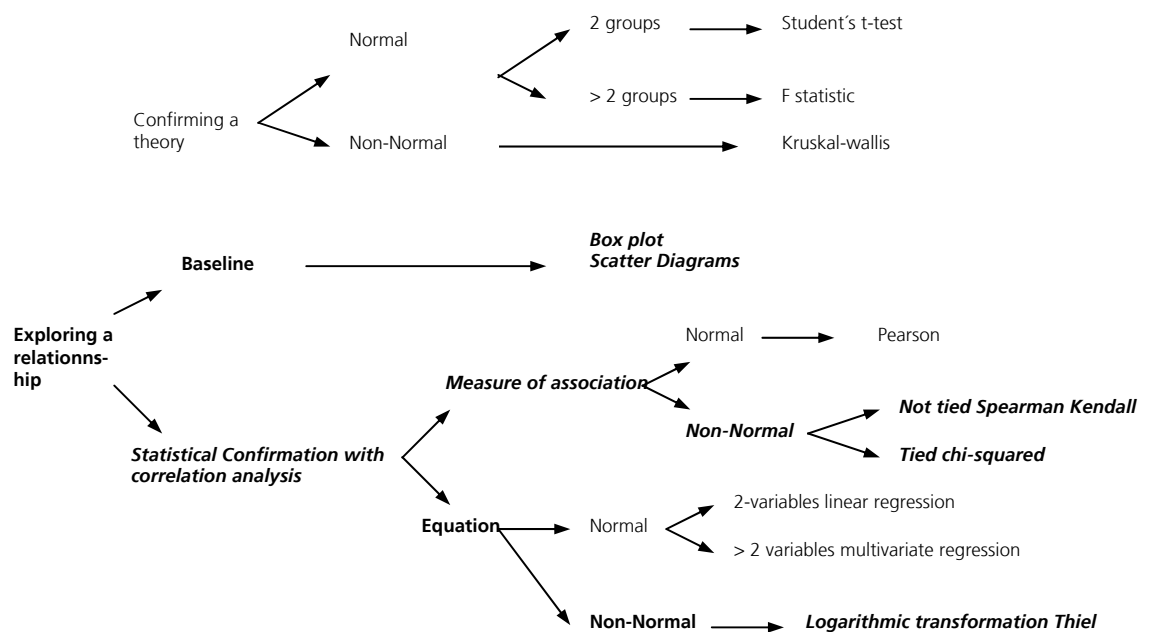


Figure 10. Decision tree for analysis techniques. [5], page 20

¹ But it must consider that non-parametric techniques have a lower statistical power than parametric ones . When a sample is normally distributed the statistical power of the non-parametric test will be less than the corresponding parametric test and as a consequence a Type II error (Selection 2.3) is more likely to be committed [10].

There are many books in statistics that might help to decide the technique to be applied [17]. Figure 3 shows the decision tree [5].

Words in ***italic & Bold*** reflect the statistical techniques in which the techniques presented in this manual are based.

The following sections present the techniques selected to analyse measures extracted from the M-System. Each of them contain:

- Description and brief introduction to the statistical technique.
- Example using the Macro.

3.3.1 Perform Descriptive Statistics & Outlier Analysis

During this activity descriptive statistics such as medians, means, inter quartile ranges and standard deviations of each measure are calculated to identify the measures. Descriptive statistics is also important to identify measures with low or zero variance. This can help explain e.g. why a measure is not useful in the later steps.

Medians and inter quartile are more appropriate than mean and variance because data set in software measurement are usually not normally distributed.

Measures of central tendency: **mean** (arithmetic average) and **median** (middlest value).

Measures of distribution: **standard deviation**, **inter quartile range** (range containing the middle 50% of data points), **minimum/maximum value** (smallest/largest value).

All these measures can be presented in a visual form called a box-plot [5]. Normally, only measures with more than five non-zero data points are considered for all subsequent analyses [1].

Outliers are data points that are located outside the upper und lower value in box plots. It is important to consider inclusion or exclusion of outliers because they can have a large influence on the analysis. Two kind of outliers are considered here [1]:

- Univariate outliers: a class that shows an extreme value in the distribution of at least one of the measures used in study. An univariate outlier is influential, if the significance of the relationship between the measure and the dependent variable depends on the absence or presence of the outlier.
- Multivariate outliers: To identify multivariate outliers in the sample-space formed by n-measures, for each data point the Mahalanobis distance (see

[1.1]) is calculated (this measure provides an indication of whether or not an observation is an outlier with respect to the independent variable to each data point calculated). **Multivariate outliers are data points with a large distance** from the sample space centroid [11]. A multivariate outlier may be over-influential if the significance of any of the variables in the model depends on the absence or presence of the outlier.

$$r^2(x) = (x - m_x)' C^{-1} (x - m_x) \quad [3.1]$$

Where:

X is the vector of raw data for the measures (independent variables)

M is the vector of means for the independent variable

C is the inverse of the matrix of deviations for the measures.

Results in descriptive statistics are important because they allow comparing with replicated studies, so researches may determine, for example, if the data collected across studies stem from similar populations [1].

Example

- 1) Click the button "Summary" to get the descriptive statistics of the data. This is the first step for the data analysis and **must be always done**. This option generates the spreadsheet (named "metrics selected") used in later options for the analysis

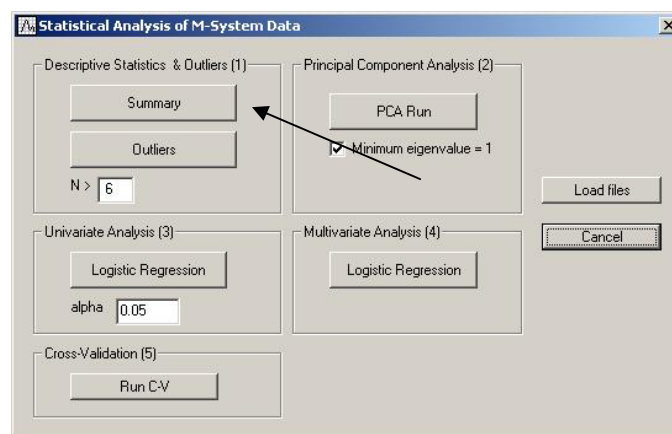


Figure 11.

Control Panel

2) Select the set of independent variables and click ok.

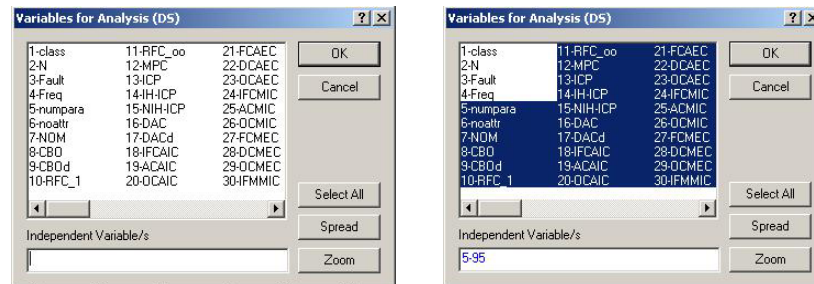


Figure 12. Dialogs for selection of variables

3) Select the dependent variable for the study and click ok.

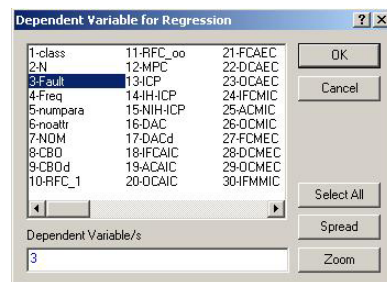


Figure 13. Dialog for selection of the independent variable

3) Select the dependent variable (counts) for the study and click ok.

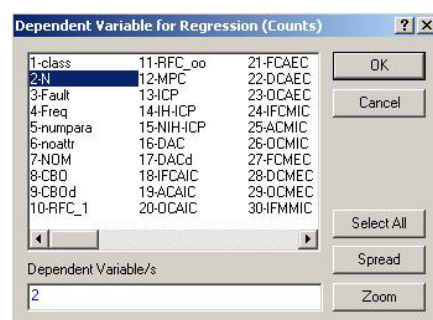


Figure 14. Dialog for selection of the independent variable (Counts)

Note that all variables involved in the data analysis must be specified; Independent variables and dependent variables (binary variable (1,0) and counts).

4) The following Workbook shows the results for Descriptive Statistics.

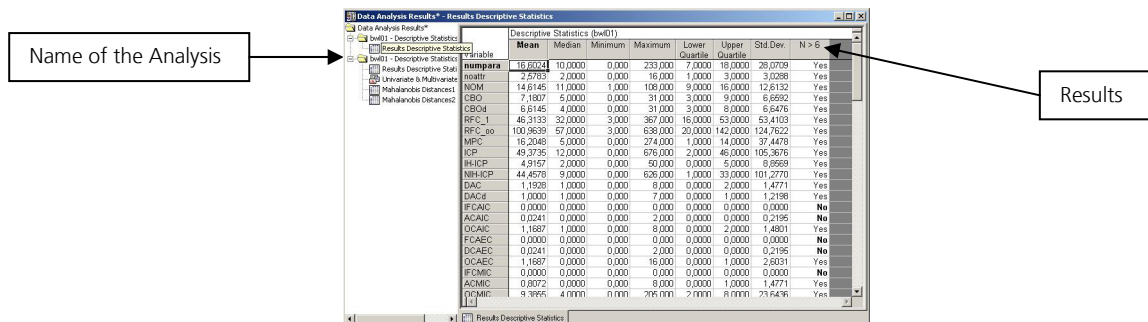


Figure 15. Results for Descriptive Statistics

The results (see figure 15) for each metric show the mean, median, minimum and maximum values, quartiles, standard deviation and Yes/No if the metric has more than N non-zero points.

By default the number of minimum non-data points to be calculated is 6, this number can be modified in the control panel.

5) Univariate & Multivariate Outliers are calculated clicking the button “Outliers”.

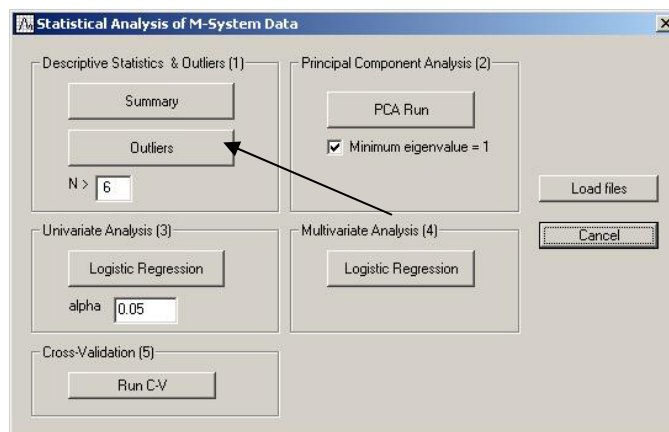


Figure 16. Outliers Option

6) Select the variables for the outliers analysis.

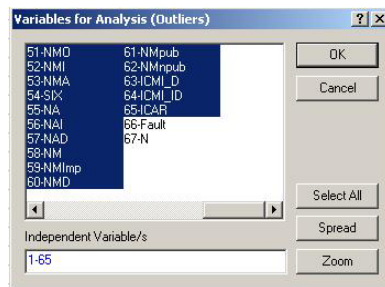


Figure 17.

Dialog for the selection of variables (Outliers)

7) The results for Univariate & Multivariate Outliers are shown in the following figures.

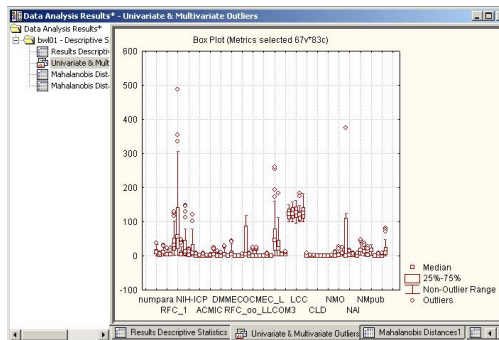


Figure 18.

Box Plots Results

Figure 18 displays the box plots for the variables selected. The variables for Univariate outliers can be chosen one by one to better visualisation.

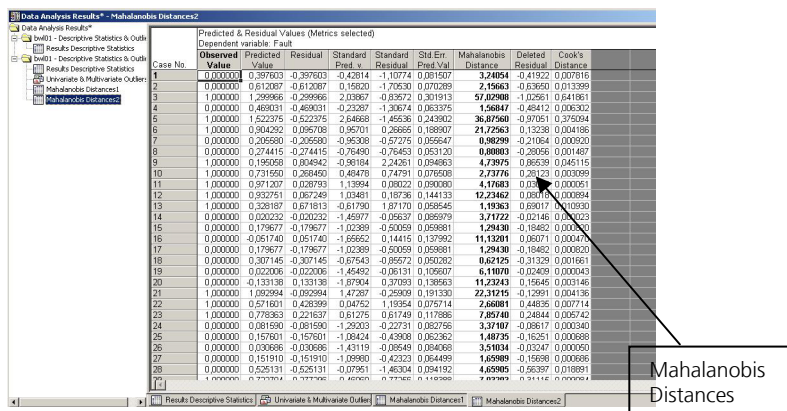


Figure 19.

Results for Outliers (Multivariate)

The screenshot shows the 'Data Analysis Results' window for 'Mahalanobis Distances'. The main table lists cases with their Mahalanobis distances. A callout box labeled 'Mahalanobis Distances' points to the 'Mahalanobis Distance' column in the table.

Case	Observed Value	Predicted Value	Residual	Standard Residual	Standard Residual	Std Err	Mahalanobis Distance
5	1.000000	1.259966	-0.259966	-0.03867	0.031513		57.80900
50	1.000000	1.113384	-0.113384	-1.52864	-0.31952	0.260166	48.97188
65	1.000000	0.940286	-0.059714	-0.03280	0.42288	0.267447	41.57289
5	1.000000	1.522375	-0.522375	-2.64668	-1.46536	0.43902	36.87560
81	1.000000	1.128694	-0.128694	-1.56679	-0.34462	0.220348	29.91561
87	1.000000	1.169795	-0.169795	-1.65275	-0.44241	0.195520	25.34369
21	1.000000	1.092894	-0.092894	-1.47287	-0.25939	0.191330	22.31215
6	1.000000	0.904292	-0.095708	-0.95701	-0.26695	0.188907	21.72563
64	1.000000	0.929731	-0.070269	-1.03666	-0.19677	0.187274	21.26716
60	1.000000	0.407879	-0.407879	-0.40005	-1.13637	0.153651	14.03870
12	1.000000	0.932751	-0.067249	-1.03481	-0.18736	0.144133	12.22462
74	1.000000	1.088333	-0.088333	-1.71245	-0.30205	0.143630	12.14700
61	1.000000	0.431631	-0.431631	-0.33639	-1.20227	0.149522	11.58105
20	1.000000	-0.131390	0.131390	-0.87804	0.27093	0.136963	11.22943
41	1.000000	0.880651	-0.119949	-0.89074	0.33418	0.130111	11.13524
16	1.000000	-0.051740	0.051740	-1.65652	0.14415	0.137902	11.13281
51	1.000000	0.494402	-0.494402	-0.16382	-1.37743	0.133378	10.33493
80	1.000000	0.968069	-0.031931	-1.13136	0.08696	0.124019	8.88171
42	1.000000	0.612284	-0.387716	-1.15874	1.08020	0.12006	8.64237
33	1.000000	0.969500	-0.030500	-1.18189	0.03736	0.122200	8.55338
68	1.000000	0.194983	-0.194983	-0.96205	-0.54323	0.120075	8.49719
52	1.000000	0.172072	-0.172072	-1.04468	-0.47940	0.120405	8.23943
63	1.000000	0.616763	-0.383237	-1.17919	1.02896	0.118966	7.86253
29	1.000000	0.722704	-0.277296	-0.46080	0.77556	0.118388	7.93293
23	1.000000	0.703363	-0.226637	-0.61275	0.61746	0.117888	7.85148
39	1.000000	0.697801	-0.002199	-1.21354	0.00613	0.116214	7.68820
40	1.000000	0.715340	-0.284660	-0.44047	0.75338	0.114845	7.48688
87	1.000000	0.494405	-0.005595	-0.16320	1.40038	0.111753	6.36892
75	1.000000	0.644260	-0.355740	-1.14702	0.00911	0.111067	6.06740

Figure 20.

Results for Outliers (Multivariate)

Figures 19 and 20 show the Mahalanobis distances of the cases, which are candidates (outliers) to be removed in later analyses. Figure 20 shows the results by distances, Figure 19 by cases. The higher the Mahalanobis distance, the more likely (the case associated) to be outlier.

For each execution of the macro, the results are store in different Workbooks. That is, only the results of same session of execution are store in the same Workbook. STATISTICA provides many utilities to copy, past, save, etc. spreadsheets.

Click the bottom "Summary", generates the spreadsheet named "metrics selected". This spreadsheets contains the remaining metrics of applying the restriction "> N non zero data points"

User may save all results in the format he/she wishes (text file, Excel file, etc).

3.3.2 Principal Component Analysis

Introduction

The measures in the data set (M-System Data) could be strongly correlated, in this case they are likely to measure the same underlying dimension (quality factors).

Principal component analysis (PCA) is a technique for analysis of multivariate data sets that provides a method for data reduction. In this case, PCA identifies the underlying and orthogonal dimensions that explain relationships among measures and it provides a technique to reduce the number of metrics with which it must deal.

PCA is applied to data set (M-System Data):

- To capture whether measures are capturing different dimensions and structural aspects. if not,
- to interpret what individual measures are really capturing and in that way
- to identify redundancy that is present among existing measures and
- to reduce the number of independent variables to a set that captures the same properties that the original ones.

Model

Principal Components Analysis is a transformation from the observed variables (measures) $x_1, \dots, x_p$ to new variables $y_1, \dots, y_p$ where

$$\begin{aligned} y_1 &= a_{11}x_1 + a_{12}x_2 + \dots + a_{1p}x_p \\ y_2 &= a_{21}x_1 + a_{22}x_2 + \dots + a_{2p}x_p \\ &\vdots \\ y_p &= a_{p1}x_1 + a_{p2}x_2 + \dots + a_{pp}x_p \end{aligned} \quad [3.2]$$

Principal components (PCs) are linear combinations of the standardized independent variables. The sum of the squares of the coefficients of the standardized variables in one linear combination is equal to one.

Principal components (PCs, $y_1, \dots, y_p$) are calculated as follows:

- The first PC (y_1) is the linear combination of all standardized variables which explain a maximum amount of variance in the data set. The second and subsequent PCs are linear combinations of all standard variables, where each new PC is orthogonal to all previously calculated PCs and captures a maximum variance under these conditions. Usually, only a subset of all variables have large coefficients- also called the loading of the variable - and therefore contribute significantly to the variance of each PC. The variables with high loadings help identify the dimension the PC is capturing but this usually requires some degree of interpretation.
- In order to identify these variables, and interpret the PCs, the **rotated components** are considered. This is a technique where the PCs are subjected to an orthogonal rotation. As a result, the rotated components show a clearer pattern of loadings, where the variables either have a very low or high loading, thus showing either a negligible or a significant impact on the PC. There exist several strategies to perform such a rotation.

Here the **varimax rotation** is used, which is the most frequently used strategy in the literature. For interpretation of the rotated components (RC), variables **with high loadings (>0.7 or <-0.7)** is a useful threshold) in each RC and identify what these measures have in common, are considered.

- For a set of n measures there are, at most, n orthogonal PCs, which are calculated in decreasing order of variance they explain in the data set. Associated with each PC is its eigenvalue, which is a measure of the variance of the PC. Usually, Only a subset of the PCs is selected for further analysis (interpretation, rotated components, etc). A typical stopping rule that can select in the macro is that only PCs whose **eigenvalue is larger than 1.0 are selected** [1] [2].

In the next section it shows how to apply PCA on the data set with the macro.

How apply PCA

- 1) Click the button "PCA Run".

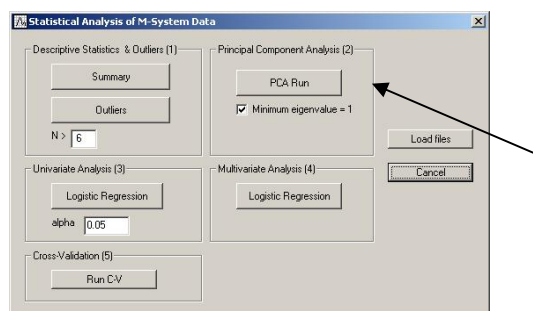


Figure 21: Control Panel. PCA Option

- 2) Select the variables for the analysis.

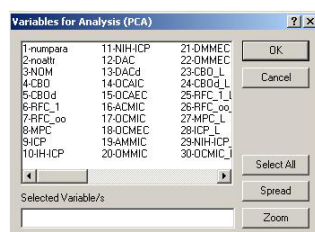


Figure 22: Dialog for the selection of variables

Results are present in the following figure.

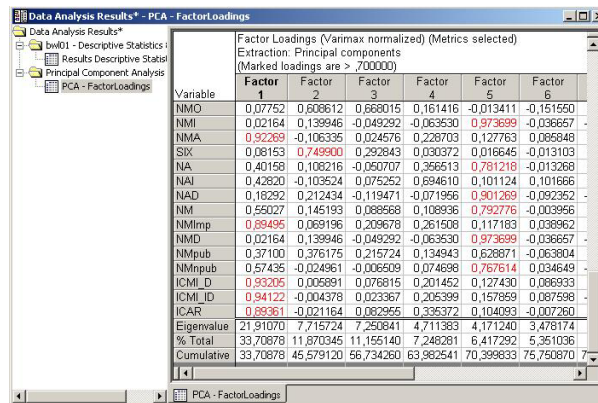


Figure 23. Results for PCA

Results are presented in figure 23. For each measure it shows its corresponding loading (see section 3.3.2 - Model). In the later rows is showed the eigenvalue of each factor, the variance of the data set explained by PCA (in percent) and the cumulative variance in the table. Loadings values higher that 0.7 are in highlight.

Restrictions: If the number of metrics to be selected if figure 22 are small, the option "minimum eigenvalue = 1" must be deselected.

3.3.3 Correlation of measures to size - Optional

The main reason why correlation of measures (coupling/cohesion etc) to size were analysed in the past, was to investigate if those measures actually measure something different than size. Size might have a great influence on the dependent variable. If size measures (that are much simpler to collect) were good predictors, than it would be preferable to use those, instead of the sophisticated coupling/cohesion measures. Nowadays, this analysis is concerned as redundant with PCA. If the size measures are included in the PCA then it can be detected if one or more dimensions are characterized by size and which measures contribute to the size dimensions.

3.3.4 Univariate Analysis

Introduction

Univariate Analysis is used to construct predictions models for each individual measure against the dependent variable to determine if the measure is statistically related. The objective is to investigate which measures have an impact on

the dependent variable. Dependent variables are the result of goal- and or hypothesis definition and should therefore be provided during the phase Quality definition; see figure 1.

The choice of a regression analysis method depends on the nature of the dependent variable at hand. Logistic and Poisson regression are useful for effort data. Both regression methods have applied by the M-System for univariate data analysis so-far [1][3]. This section focuses on Logistic regression mainly, because Poisson has the advantage of an unbiased effort prediction, but is also a very specific approach.

Logistic Regression

Univariate logistic regression is a statistical standard technique that applies the following equation:

$$\pi(x) = \frac{e^{(\beta_0 + \beta_1 x)}}{1 + e^{(\beta_0 + \beta_1 x)}} \quad [3.3]$$

- π (*conditional probability*) is the probability that an event occurs, taking into account that the objective is to assess the impact of each measure on the dependent variable.
- x indicates the design measure used as independent variable in the model (also called covariate of the logistic regression equation)..
- The coefficients β_0 and β_1 are estimated through the maximization of likelihood functions (maximum likelihood estimation)
- The curve between π and x_i (*in the multivariate case*)– assuming that all other x_j – takes a flexible S shape which ranges between two extreme cases:
 - 1) When a variable is not significant, then the curve approximates a horizontal line, i.e., π does not depend on x_i .
 - 2) When a variable entirely differentiates the dependent variable software parts, then the curve approximates a step function.
- The coefficients β_i are estimated through the maximization of a likelihood function, built in the usual fashion, i.e., as the product of the probabilities of the single observations, which are functions of the covariates (whose values are known in the observations) and the coefficients (which are unknown). For mathematical conveniences, $l = \ln[L]$, the loglikelihood, is usually the function to be maximized. This procedure assumes that all observations are statistically independent.
- To assess the impact of each measure on the dependent variable, the odds ratio is used, because the regression coefficients β_i cannot be easily inter-

preted for this purpose. [9]. The Odds ratio $\psi(X)$ represents the ratio between the probability of the event in the dependent variables occurs and the probability of not, when the value of the measure is X.

$$\Delta\psi = \frac{\psi(X + \sigma)}{\psi(X)}, \text{ where } \psi(X) = \frac{\pi(X)}{1 - \pi(X)} \quad [3.4]$$

where σ is the standard deviation of the measure. Therefore, $\Delta\psi$ represents the reduction/increase in the odds ration when the value X increases by one standard deviation.

- To asses the statistical significance of each independent variable in the model, a likelihood ratio chi-square test is used.
- Let $l = \ln [L]$ be the loglikelihood of the model given in equation 3.3, and l_i be the loglikelihood of the models without variable x_i . Assuming the null hypothesis that the true coefficient of X_i is zero, the statistic $g = -2(l - l_i)$ follows a chi-square distribution with one degree of freedom ($\chi^2(1)$). Therefore, $p\text{-value} = P(\chi^2(p) > G)$ it will be tested. If p is larger than some level of significance α , the observed change in the loglikelihood may be due to chance, and X_i is not considered significance. If $p \leq \alpha$, X_i , the observed change in the loglikelihood in unlikely to be due to chance, and x_i is considered significant.

Assessing Fit

To fit the model the following formula Logistic regression R^2 is used (not be consused with the least-square regression R^2). The higher R^2 , the higher the effect of the modes's explanatory variables, the more accurate the model. However, as opposed to the R^2 of least square regression, high R^2 s are rare in logistic regression:

R^2 is defined by the following ration.

$$R^2 = \frac{LL_s - LL}{LL_s}, \quad [3.5]$$

where:

LL is the loglikelihood obtained by Maximum Likkelihood Estimation
 LL_s is the loglikelihood obtained by Maximum Lilelihood Estimation of a model without any variables, i.e., with only β_0

Details about Logistic Regression can be found in [9].

Poisson Regression – Optional step

Poisson Regression is applied to the analysis of counts of events. In this case (M-System data), the dependent variable y is assumed to have Poisson distribution with parameter μ .

$$\Pr(y | x_1, \dots, x_n) = \frac{e^{-\hat{\mu}} \hat{\mu}^y}{y!}, \quad \hat{\mu} = e^{\beta_0 + \beta_1 x_1 + \dots + \beta_n x_n}$$

where the parameter μ is the expected value and variance of y : $\mu = E(y) = \text{Var}(y)$ (equidispersion).

For a given data set, the regression coefficients $\beta_0, \dots, \beta_n$ can be estimated through maximization of the likelihood function.

But for the data analysis, the negative binomial model is used instead. The negative binomial model is an extension of the Poisson distribution, which allows the variance of the process to differ from the mean. That is, $\text{Var}(y | X) > E(y | X)$, known as overdispersion, what is more commonly found in practice [3].

$$\Pr(y | x_1, \dots, x_n) = \frac{\Gamma(y + v)}{y! \Gamma(v)} \left(\frac{v}{v + \hat{\mu}} \right)^v \left(\frac{\hat{\mu}}{v + \hat{\mu}} \right)^y, \quad \hat{\mu} = e^{\beta_0 + \beta_1 x_1 + \dots + \beta_n x_n}$$

with $v > 0$ that is estimated along with the regression coefficients β_0, β_1 , in a maximum likelihood estimation based on the above probability distribution.

$$\hat{\mu} = E(y | X) \text{ but } \text{Var}(y | X) = \hat{\mu} + \alpha \hat{\mu}, \text{ where } \alpha = v^{-1}$$

α is called the dispersion parameter. For $\alpha \rightarrow 0$, the negative binomial model converges towards the Poisson model.

According to all these assumptions, the next two sections show two examples with both techniques.

Details about Poisson Regression can be found in [17]

Example: Logistic Regression

1) Click the button "Logistic Regression" from Univariate Analysis.

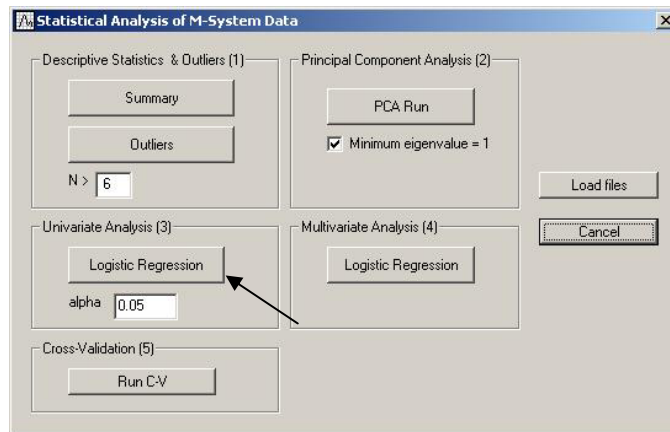


Figure 24. Control Panel. Logistic Regression.

2) Select the variables for the analysis. Dependent variable (left), independent variables (right).

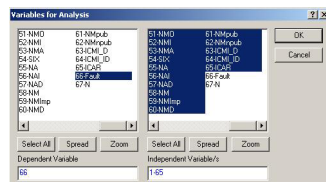


Figure 25. Dialogs for the selection of variables

Click Ok.

Note that the selection of the variables cannot be overlap.

3) The results for Logistic Regression are presented in the following figure.

	1	2	3	4	5
	Coeff	Std Err	p-value	R ²	Odds Ratio
humpara	0.036654	0.023028	0.034617	0.039128	1.00082136
humparr	0.26773	0.124037	0.00388	0.073097	1.03633376
humparab	0.003296	0.039662	0.00586	0.006664	1.00382984
CBO	0.265885	0.085339	1.004E-6	0.188882	1.02469725
CBOd	0.321006	0.092486	5.537E-7	0.219862	1.03012712
RFC_L	0.011995	0.008027	0.033176	0.038889	1.00006339
RFC_os	0.002837	0.002064	0.142263	0.018876	1.00000585
MPC	0.073918	0.030336	0.000139	0.127211	1.00224493
ICP	0.040081	0.015962	1.00E-6	0.200434	1.00001617
NIH-ICP	0.040749	0.030329	0.144372	0.018677	1.00123663
NIH-ICP	0.059353	0.019050	3.141E-7	0.229312	1.00113170
DAC	0.003	0.190491	0.964093	0.404E-6	1.00046155
DACd	-0.03305	0.181596	0.955667	0.00029	0.98401589
OCAC	-0.01706	0.14876	0.903369	0.000114	0.99744861
OCACd	0.113651	0.303472	0.000733	0.099969	1.24092743
ACMIC	0.323196	0.19068	0.060434	0.033902	1.06366553
OCMIC	0.010772	0.014418	0.363727	0.007291	1.00016532
OCMEC	0.093261	0.044008	0.000467	0.107675	1.00395996
AMMIC	0.145537	0.088871	0.044937	0.03246	1.01301799
OMMIC	0.073911	0.033094	0.000278	0.116837	1.0023749
OMMEC	0.068893	0.063552	0.092251	0.031437	1.00886865
OMMEC	0.045473	0.019799	0.001672	0.087184	1.00900075
CBO_L	1.734563	0.513216	0.000111	0.130895	2.43662947
CBOd_L	1.734563	0.513216	0.000111	0.130895	2.43662947
RFC_L_L	0.45226	0.158402	0.000029	0.163	1.00463235
RFC_os_L	0.010289	0.006795	0.066674	0.024479	1.00006904
MPC_L	0.055478	0.250152	2.928E-6	0.191639	1.17817789
ICP_L	0.244798	0.050396	2.106E-6	0.197186	1.02295116
NIH-ICP_L	0.244798	0.050396	2.106E-6	0.197186	1.02295116

Figure 26.

Results for Univariate Logistic Regression.

For each metric is presented the results of applying the equation [3.3] to each metric: its coefficient in the equation (3.3), Standard deviation, p-value, R² and Odds Ratio like is defined in this Section (3.3.4). Significant p-values are highlighted.

3.3.5 Multivariate Analysis.

Introduction

Multivariate Analysis is used to investigate the relationship between one dependent variable and two or more independent variables (measures). This analysis is conducted to determine how well multiple measures predict the dependent variable, when the measures are used in combination.

Logistic Regression

In Multivariate Logistic Regression, the same assumptions as in Univariate Logistic Regression (see section 3.4.4) are taken into account. In this case (two or more variables) the equation remains:

$$\pi(x_1, \dots, x_n) = \frac{e^{(\beta_0 + \beta_1 x_1 + \dots + \beta_n x_n)}}{1 + e^{(\beta_0 + \beta_1 x_1 + \dots + \beta_n x_n)}} \quad [3.6]$$

Where coefficients, statistical significance and the assessing fit of model are calculated in the same way like in 3.4.4.

Sometimes the number of measures (number of independent variables) are too large to be treated. To reduce them and select the measures to be used in the model, a strategy must be employed that

- Minimizes the number of independent variables in the model. Using too many independent variables can have the effect of increasing the estimated standard error of the model's prediction, making the models more dependent on the data set, i.e., less generalizable. A rule of thumb is to have at least ten data points per independent variable in the model.
- Reduces multicollinearity, i.e., independent variables which are highly correlated. This makes the model more interpretable. Multicollinearity reduction is performed calculating the conditional number [1], a value which is based in the correlation matrix of the independent variables (measures) in the model. According to [9], tests for multicollinearity used in least-squares regression are also applicable in the context of logistic regression. Here is applied the conditional number [19 –belsley]), which is based on the conditional number of the correlation matrix of the covariates in the models. The conditional number is defined as:

$$\lambda = \sqrt{\frac{l_{\max}}{l_{\min}}},$$

where

$l_{\max}$ is the largest eigenvalue of the principal components.

$l_{\min}$ is the smallest eigenvalue.

A large conditional number indicates the presence of multicollinearity. The degree of multicollinearity is harmful, and corrective actions should be taken, when the conditional number exceeds 30 [1].

To build multivariate models (regardless of logistic/poisson/OLS regression), a stepwise selection process can be used, which subsequently adds/removes variables to a model until its goodness of fit can not be improved anymore. The two major stepwise selection processes used in logistic regression are forward selection and backward elimination. The general forward selection procedure starts with a model that includes the intercept only. Based on certain statistical criteria, variables are selected one at a time for inclusion in the model, until a stopping criteria is fulfilled. Because of the large number of independent variables (M-System Data), the **forward selection** procedure to build the prediction models is selected. In each step, all variables not already in the model are tested: the most significant variable is selected for inclusion in the model. If this causes a variable already in the model to become not significant (at $\alpha_{\text{exit}}=0.10$), it is deleted from the model. The process stops when adding the best variable no longer improves the model significantly (at $\alpha_{\text{enter}} = 0.05$). The significance of a variable is tested by a loglikelihood ratio test (see section 3.4.4)

Example: Logistic Regression

1) Click the button "Logistic Regression" from Multivariate Analysis.

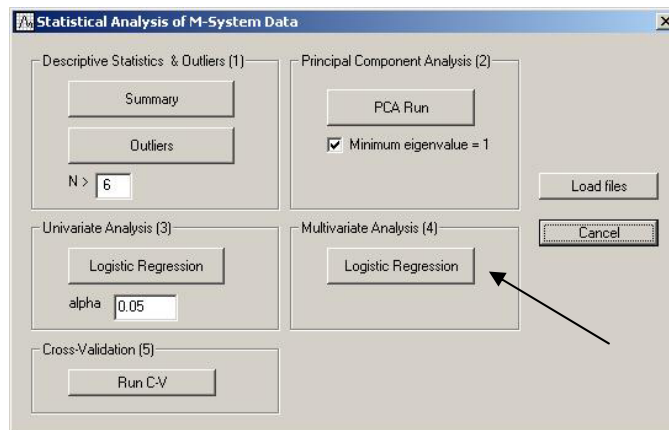


Figure 27.

Control Panel.

2) Select the variables involved in the analysis. Note that it must perform some kind of interpretation on the previous results to carry out this one

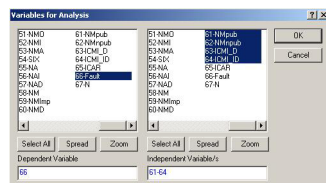


Figure 28.

Dialogs for the selection of variables

Click Ok.

The results are presented in the following figures.

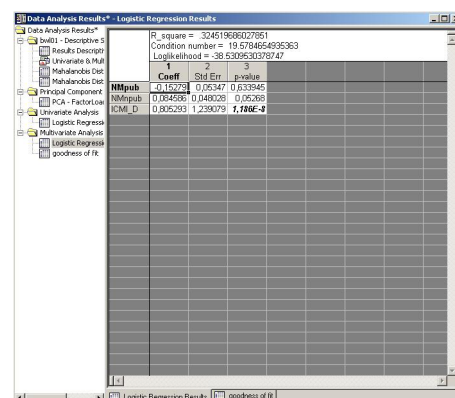


Figure 29.

Results for the model

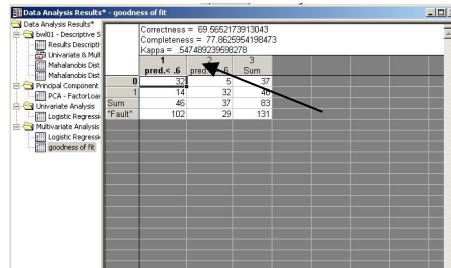


Figure 30. Goodness of fit of the model

Two spreadsheets are generated for Multivariate analysis. One (figure 29) shows the results for the model (equation 3.6): Coefficient, Std. Deviation a p-value. The heading in the spreadsheet shows the values for R square, Conditional Number and Loglikelihood for the model generated. The second Spreadsheet (figure 30) shows the results of goodness of fit for the model. Completeness, Correctness and Kappa- value like is defined in the next section (3.3.6)

3.3.6 Evaluating estimation accuracy

The accuracy of the prediction model (that describes the relationships between independent and dependent variables) should be determined to provide information about the probability that the data analysis outcomes are right.

Several techniques to the determine the accuracy of the prediction models exist: Magnitude of Relative Error (MRE), Completeness, Correctness, Kappa. The type of regression techniques and the dependent variables will determine the suitability of the accuracy technique.

The following measures might be used to evaluate the goodness of fit of the prediction models:

- Completeness: the degree to which all the parts of a software system or component are present and each of its parts is fully specified and developed [12]. Here, the completeness is concerned with which % of the classes fulfils a specific criteria.
- Correctness: the degree to which a system or component is free from faults in its specification, design and implementation. [13]. In this case the completeness in concerned with % of the classes classified correctly (according to a criteria).
- MRE (Magnitude of Relative Error) or ARE (Absolute Relative Error).

$$MRE = \frac{ARE}{eff_i}, \text{ where } ARE = |eff_i - \hat{eff}_i|$$

- R^2 (see section 3.4.4) is also used as a measure of the goodness of fit. But this measure is specific for Logistic Regression techniques based on maximum likelihood estimation.
- Kappa [14] is a measure of the degree of agreement of two variables. Appendix B elaborates on Kappa.

3.3.7 Cross-Validation

Introduction

The results of accuracy of the prediction models might be optimistic because of they are applied to the same data set which is derived from. These prediction models should be applied to a different data set.

Cross-validation is a common method used for model checking in regression problems that will help in this task.

The steps for making a Cross-validation are described as follows:

- Divide the data set into k subsets of similar size (normally $k = 10$).
- Each time, one of the k subsets is used as the test set, and the other $k-1$ subsets are put together to form a training set. It re-fits the model using the $k-1$ subsets and then applies the model to the currently partition (k subset). The point is to re-apply the techniques presented in Section 3.3.6 to newly predicted values.

Examples: Cross-Validation

1) Click the button "Cross-Validation"

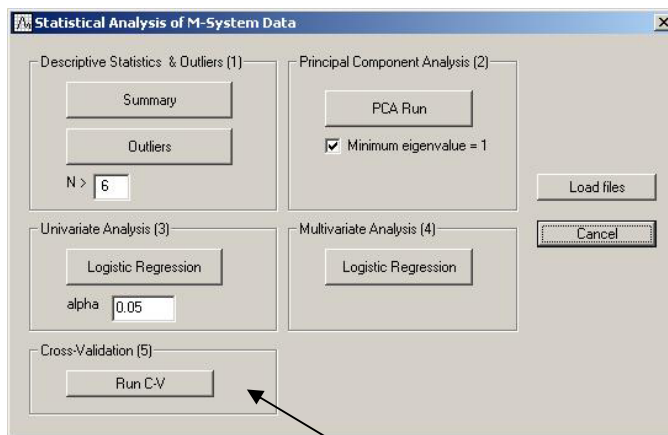


Figure 31.

Control Panel.

2) Select the variables involved in the analysis. Note that It must perform some kind of interpretation on the previous results to carry out this one

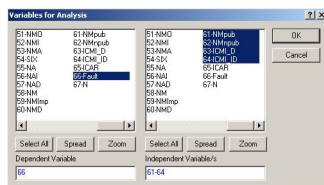


Figure 32.

Dialog for selection of variables

The results are provided in the following figure.

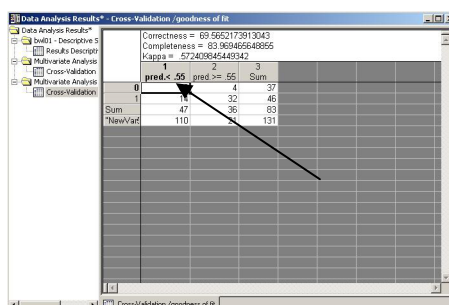


Figure 33.

Results for Cross-Validation.

To select the threshold (see figure 33) for the cross-validation, the percentage of classes being classified fault-prone is roughly the same as the percentage of classes the actually are fault-prone, so that the threshold is selected to roughly balance the number of actual and predicted fault-prone classes.

3.3.8 Model Application

The prediction models built in the last section (3.3.7) may be applied in practice, the results obtained from the predictions models can be presented in plots, (for example the classes can be sorted in decreasing order of their predicted fault-proneness), so that when planning inspections during the software development process, it can be able to make a trade-off between the resources spent on inspections and effectiveness as well as the evaluation of its performance.

4 Summary

The main goal of this manual was to describe the process of analysing M-System Data (Software Measurement data) using statistical techniques to investigate the relationships between measures and quality attributes (fault-proneness, effort, etc).

This process implies to state the definition of the goals, variables of interests and hypotheses to be tested.

To analyse the data, a first study about the nature and distribution of the data can draw conclusions that could to address the analysis to different techniques.

Descriptive Analysis and Outliers is a first step for this analysis. The great amount of metrics measured with M-System could be too much information to be analysed with statistical techniques. PCA (Principal Component Analysis) provide a method to reduce the number of metrics to smaller amount of them, in which it is represented the same information.

Univariate and Multivariate Logistic regression are applied to analyse the relationships between the measures extracted and the dependent variable. Poisson or Logistic Regression are applied according to criteria.

Finally all models developed with the statistical techniques are validated.

A Tool that supports the process presented has been developed. The following table summaries the process.

What do you want to do?	Number of variables	Statistical technique	
Distribution of data		Boxplots Descriptive Statistics	
Investigate the correlation with Size		Optional ----- PCA	
Explore the relationships among variables	One independent variable	Univariate logistic regression	Logistic Regression
			Poisson
	> two independent variables	Multivariate logistic regression	Logistic Regression
			Poisson

Figure 34.

Summary

5 References

- [1] L.Briand, J.Wüst, W. Daly, V.Porter, "Exploring the Relationships between Design Measures and Software Quality in Object-Oriented Systems".
- [2] L.Briand, J.Wüst, H. Lounis, "Investigating Quality Factors in Object-Oriented Designs: An Industrial Case Study". ISERN 98/29
- [3] L.Briand, J.Wüst "The Impact of Design Properties on Development Cost in Object-Oriented Systems. ISERN 99/16
- [4] J.Wüst, T. Punter "M-System User Manual", IESE Report 37.01/E
- [5] N.E. Fenton, S.L.Pfleeger, „Software Metrics: A Rigorous & Practical Approach". International Thomson Computer Press, Second Edition, 1996
- [7] F.Fioravanti, P. Nesi, "A study on fault-proneness detection of Object-Oriented systems".
- [8] ,Stated of the Art Report in Empirical Studies.
- [9] D. Hosmer, S. Lemeshow "Applied Logistic Regression", John Wiley & Sons, 1989.
- [10] L. Briand, J.Daly, P.Fusaro, Design of empirical studies. Fhg IESE, 1996
- [11] L. Briand, E. Arisholm, S. Counsell, F. Houdek, P. Thévenod-Fosse, "Empirical Studies of Object-Oriented Artifacts, Methods and Processes: State of the Art and Future Directions". Technical Report ISERN-99-12.
- [12] Boehm, Barry W.; Brown, John R.; Kaspar, Hans; Lipow, Myron; MacLeod, Gordon J. & Merritt, Michael J. Characteristics of Software Quality. New York, NY: North-Holland Publishing Company, 1978.
- [13] Institute of Electrical and Electronics Engineers. IEEE Standard Computer Dictionary: A Compilation of IEEE Standard Computer Glossaries. New York, NY: 1990.
- [14] Cohen, J., 1960. A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement, Vol. XX, No.1

- [15] Landis, J. & Koch, G. G. (1977). The measurement of observer agreement for categorical data. Biometrics, (33), 159-174.
- [16] Sheskin, D.J. „Handbook of Parametric and Nonparametric statistical Procedures“. Second Edition. CHAPMAN & HALL/CRC.
- [17] S. Long, “Regression Models for Categorical and Limited Dependent Variables”, Advanced Quantitative Techniques, Sage publications, 1997.
- [18] Duntelman, G., 1989. Principal Component Analysis. Sage University PAper 07-69, Thousand Oaks, CA.

Statistical Tools

- [18] STATISTICA (<http://www.statsoftinc.com>)
- [19] S-plus (<http://www.insightful.com>)
- [20] R (<http://www.r-project.org>)
- [21] Statistics resources (<http://www.statistics.com>)

6 Appendix A.1: Tool Usage

Hardware & Software Requirement

The Macro used for the statistical analysis has been implemented under STATISTICA. STATISTICA 6.1 of StatSoft ® [18] should be therefore installed in your computer before starting with the analysis.

7 Appendix A.2: Statistica Macro

The following describes the installation of the Statistica macro.

- 1) Open STATISTICA Application
- 2) Go to Open in menu File or click the open button in the navigation menu.
- 3) Load the file "Macro for statistical Analysis of M-System data.svb"
- 4) To execute the macro, click in the

How to Start

- 1) Initialise the STATISTICA environment.

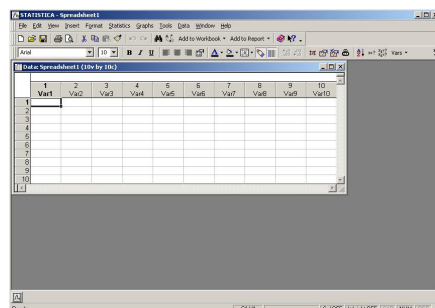


Figure 35. STATISTICA environment

- 2) Go to Tools-Macro-Macros.

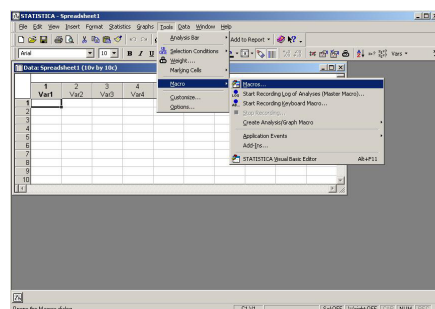


Figure 36. Options for load a macro.

3) Open or **Run** the macro called “**Macro for Statistical Data Analysis of Code Measurement Data.svb**”

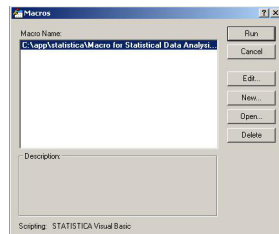


Figure 37.

Macro form

The source code of the macro is visualised (if open).

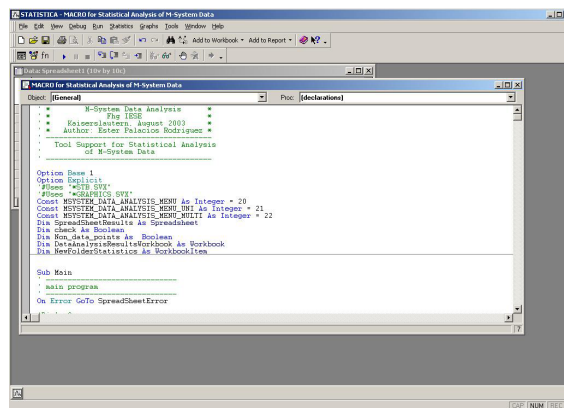


Figure 38.

Source Code

3) To execute the application, please click F5, the icon , or go to Tools-Macro-Macros and click Run.

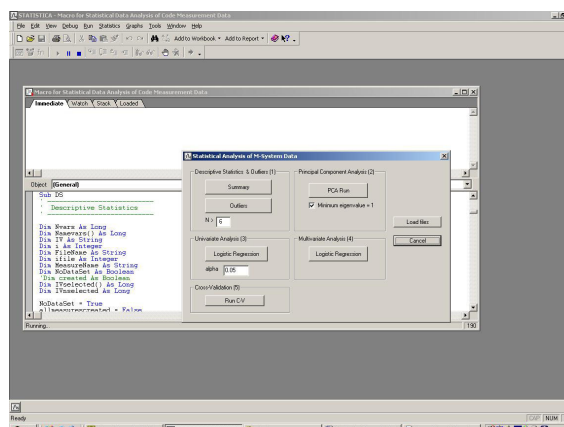


Figure 39.

Application form

Section 3.1 explains how to read the files.

8 Appendix B: Kappa

Kappa [14] is a measure of the degree of agreement of two variables. Kappa are based in the following formula.

$$K = \frac{P_0 - P_e}{1 - P_e} \quad \text{or} \quad K = \frac{\frac{1}{n} \sum_{i=1}^r n_{ii} - \frac{1}{n^2} \sum_{i=1}^r n_i n_i}{1 - \frac{1}{n^2} \sum_{i=1}^r n_i n_i}$$

Figure 39.

Kappa coefficient.

Where:

- $P_0 = \frac{1}{n} \sum_{i=1}^r n_{ii}$: Level of agreement before prediction
- $P_e = \frac{1}{n^2} \sum_{i=1}^r n_i n_i$: Level of agreement after prediction

		Variable A						
		1	2	3	...	r		Total
Variable B	1	n_{11}	n_{12}	n_{13}	...	n_{1r}		n_1
	2	n_{21}	n_{22}	n_{23}	...	n_{2r}		n_2
	...				...			
	r	n_{r1}	n_{r2}	n_{r3}	...	n_{rr}		n_r
	Total	n_1	n_2	n_3	...	n_r		N

Interpretation of k-values [15]:

K-values	Interpretation
Below 0.00	Poor
0.00-0.20	Slight
0.21-0.40	Fair
0.41-0.60	Moderate
0.61-0.80	Substantial
0.81-1.00	Almost perfect

Kappa values (k) range between –1 and 1, the higher the value, the better the agreement. A Kappa zero indicates that the agreement is no better than what can be expected from chance.

Example:

		predicted		Σ
		$\pi \leq 0.75$	$\pi > 0.75$	
Actual	No fault	49 classes	31 classes	80 classes
	Fault	31 classes	15 classes	46 classes
Σ		80 classes	46 classes	126 classes

Taking into account the table above, Kappa coefficient remains:

$$P_0 = \frac{1}{126} (49 + 15) = 0,507$$

$$P_e = \frac{1}{126^2} (80 * 80 + 46 * 46) = 0,536$$

$$k = \frac{0,507 - 0,536}{1 - 0,536} = -0,625$$

The Kappa-value results is –0,625. According to the interpretation of k-values, the results show a substantial agreement of the variables.

9 Appendix C: Validity threats

It is important to consider the validity of the results. According to [8], results can be validated from four points of view:

- Conclusion validity (fig.8, **1**): is concerned with the way to draw the correct conclusion. In this case, the selection of a particular tool to extract the data set (M-System) and process of analysis selected (see fig. 2).

The fact of select cases of study with small sample size or make some assumptions about the distribution of data to choose the “correct” statistical test, can entails results with a high margin of error or conclusions which are difficult to be generalized.

- Internal validity (fig.8, **2**): The dependent variables may not only be influenced by the independent variables. Internal validity is concerned with the degree to which conclusions can be drawn about the causal effect of the independent variables on the dependent variables [1].

Investigate whether the relationship between the measures and the dependent variable is due to causal effects or not is a risk to be studied.

- Construct validity (fig.8, **3**): The selection of the correct independent and dependent variables (see Section 2.3) can also influence the conclusions. The selected variables could not measure accurately the factors that they try to measure.

- External validity (fig.8, **4**): The degree to which the results of the research can be generalized to the population under study and other research setting [1].

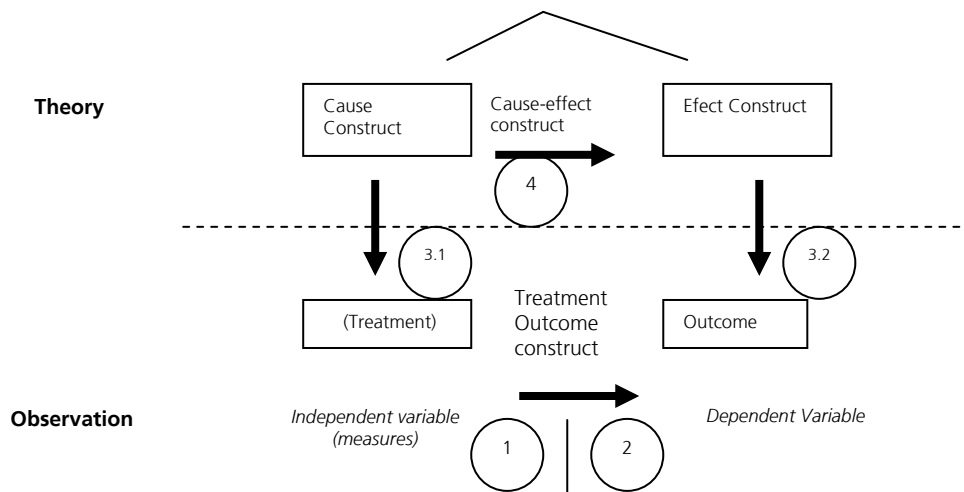


Figure 40. types of validity [8] .

Index

- α , 10
- . Process of analysing for M-SystemData, 4
- Alternative Hypothesis H_1 , 10
- box-plot, 12
- Decision tree, 11
- dependent variables, 9
- Descriptive statistics, 12
- Hypotheses, 8
- independent variables, 9
- inter quartile range, 12
- Macro for Statistical Analysis of M-System Data, 6
- mean, 12
- median, 12
- M-System, 1
- M-system process, 1
- Null Hypothesis H_0 , 9
- standard deviation, 12
- STATISTICA [17], 2
- Statistical Inference, 9
- statistical significance, 10
- Type I error, 10
- Type II error, 10
- variable, 9

Document Information

Title: A Handbook for Statistical
Data Analysis of Code
Measurement Data

Date: September 30, 2003
Report: IESE-140.03/E
Status: Final
Distribution: Public

Copyright 2003, Fraunhofer IESE.
All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means including, without limitation, photocopying, recording, or otherwise, without the prior written permission of the publisher. Written permission is not needed if this publication is distributed for non-commercial purposes.